EUROPEAN JOURNAL OF EMERGING ENGINEERING AND MATHEMATICS

Adversarial Robustness In Time Series And Computer Vision Models: A Unified Theoretical And Empirical Examination Of Attacks, Defenses, And Methodological Frontiers

Prof. Oliver Schmidt
Heidelberg University, Germany

Prof. James Anderson
University of Melbourne, Australia

V0LUME 02 ISSUE 02 (2025)
Published Date: 07 August 2025 // Page no.: - 6-11

ABSTRACT

The accelerating integration of deep learning models into safety-critical, economic, and societal decision-making systems has foregrounded the urgent challenge of adversarial robustness. While early adversarial machine learning research was predominantly anchored in computer vision, recent advances demonstrate that time series models—widely deployed in finance, healthcare, cybersecurity, climate forecasting, and industrial monitoring—are equally vulnerable to carefully crafted perturbations. This article presents a comprehensive, theory-driven, and empirically grounded examination of adversarial attacks and defenses across both computer vision and time series learning paradigms, with a particular emphasis on neural architectures, evaluation protocols, and ensemble-based strategies. Drawing upon an extensive and diversified body of literature, the study synthesizes foundational adversarial concepts, modern attack mechanisms, and contemporary defense methodologies, including adversarial training, distillation, ensemble learning, and robustness-oriented architectural design. The analysis is guided by the recognition that adversarial vulnerability is not an incidental artifact of specific models, but rather an emergent property of high-dimensional statistical learning systems optimized under conventional empirical risk minimization. By integrating insights from seminal theoretical works and recent empirical contributions, including comprehensive surveys and domain-specific studies, this article articulates a unified conceptual framework that bridges computer vision and time series classification. The methodological section outlines a rigorous literature-driven analytical approach that emphasizes interpretive synthesis over numerical benchmarking, enabling cross-domain comparison without reliance on visual or mathematical formalism. The results section presents an in-depth descriptive analysis of adversarial behaviors, highlighting recurring patterns of susceptibility, transferability, and defense trade-offs observed across domains. The discussion extends these findings into a broader scholarly debate, interrogating assumptions about robustness, the limits of current defenses, and the epistemic implications of adversarial machine learning for trustworthy artificial intelligence. The article concludes by identifying critical research gaps and proposing theoretically informed directions for future work, particularly in the development of domain-aware defenses and evaluation frameworks that move beyond narrow threat models.

Keywords: Adversarial machine learning; Time series classification; Neural network robustness; Ensemble defenses; Computer vision security; Deep learning reliability.

INTRODUCTION

The rapid proliferation of deep learning systems across diverse application domains has reshaped contemporary data-driven decision-making, enabling unprecedented performance in perception, prediction, and control tasks. From image recognition and autonomous navigation to financial forecasting and medical diagnosis, neural networks have become foundational components of modern intelligent systems. However, alongside these advances, a parallel body of research has revealed a profound and unsettling vulnerability: the susceptibility of learned models to adversarial perturbations—carefully designed inputs that induce erroneous predictions while remaining imperceptible or semantically insignificant to human observers (Goodfellow et al., 2014; Szegedy et al., 2013). This discovery has catalyzed an expansive research agenda focused on understanding, exploiting, and mitigating adversarial behavior in machine learning systems, particularly within the domain of computer vision (Akhtar & Mian, 2018; Carlini & Wagner, 2017).

Early investigations into adversarial examples framed the phenomenon as a peculiarity of deep neural networks trained for image classification, emphasizing gradient-based attacks and pixel-level perturbations (Goodfellow et al., 2014). Subsequent studies, however, demonstrated that adversarial vulnerability is neither confined to specific architectures nor limited to visual data, but instead reflects deeper statistical and geometric properties of high-dimensional learning systems (Madry et al., 2017; Papernot et al., 2016). As a result, adversarial machine learning has evolved into a broad interdisciplinary field,

encompassing theoretical analysis, attack taxonomy, defense mechanisms, and domain-specific adaptations (Joseph et al., 2018; Pitropakis et al., 2019).

In parallel with developments in computer vision, time series analysis has undergone a methodological transformation driven by deep learning. Neural architectures such as convolutional networks, residual networks, and inception-style models have demonstrated state-of-the-art performance on benchmark datasets and real-world temporal tasks, often surpassing traditional statistical approaches rooted in classical time series theory (Ismail Fawaz et al., 2020; He et al., 2016). Public repositories such as the UCR Time Series Archive have played a pivotal role in standardizing evaluation and accelerating progress in time series classification research (Dau et al., 2019). Yet, despite the growing reliance on deep learning for temporal data, systematic investigation of adversarial robustness in this domain has lagged behind computer vision, creating a critical gap in both theory and practice (Galib & Bashyal, 2023).

Recent contributions have begun to address this imbalance by adapting adversarial attack methodologies to time series models and by demonstrating that temporal classifiers are highly vulnerable to both white-box and black-box perturbations (Ding et al., 2023; Dong et al., 2023). These studies reveal that adversarial manipulation of time series data can exploit temporal dependencies, frequency characteristics, and model-specific decision boundaries in ways that parallel—but also fundamentally differ from—image-based attacks. Comprehensive surveys have further emphasized that adversarial threats in time series learning demand distinct analytical frameworks, given the structured, sequential, and often context-dependent nature of temporal data (Akhtar et al., 2021; Galib & Bashyal, 2023).

Theoretical interpretations of adversarial vulnerability have generated substantial scholarly debate. One line of argument attributes adversarial examples to the linear behavior of neural networks in high-dimensional spaces, suggesting that small but systematically aligned perturbations can accumulate to produce large changes in model output (Goodfellow et al., 2014). Another perspective emphasizes the role of data distribution mismatch and insufficient coverage of adversarial regions during training, leading to brittle decision boundaries that fail under worst-case inputs (Madry et al., 2017). These interpretations have informed the development of defenses such as adversarial training, defensive distillation, and ensemble-based approaches, each of which seeks to improve robustness through different mechanisms and with varying trade-offs (Papernot et al., 2016; Deng & Mu, 2023).

Despite significant progress, the literature remains fragmented along domain lines, with computer vision and time series research often advancing in relative isolation. Surveys focused on vision-centric attacks and defenses provide limited insight into temporal models, while time series–specific studies frequently lack integration with broader adversarial theory (Akhtar et al., 2021; Machado et al., 2021). This fragmentation obscures shared principles and inhibits the transfer of robust methodologies across domains. Moreover, evaluation practices vary widely, complicating meaningful comparison of robustness claims and reinforcing the need for unified conceptual frameworks (Carlini & Wagner, 2017; Apruzzese et al., 2023).

The present article addresses these challenges by offering an extensive, integrative analysis of adversarial attacks and defenses spanning computer vision and time series learning. Rather than proposing a new algorithm or benchmark, the study adopts a synthesis-oriented research design that emphasizes theoretical depth, historical context, and critical comparison of scholarly perspectives. By systematically examining how adversarial concepts manifest across data modalities, architectures, and application contexts, the article seeks to illuminate underlying commonalities while respecting domain-specific nuances (Akhtar et al., 2021; Ding et al., 2023).

A central motivation for this work lies in the growing deployment of time series models in adversarially exposed environments. Financial markets, for instance, are influenced by strategic actors capable of manipulating signals and information flows, rendering predictive models vulnerable to adversarial exploitation (Corizzo & Rosen, 2024). Similarly, cybersecurity systems based on temporal network traffic patterns face adaptive adversaries who actively probe and evade detection mechanisms (Rosenberg et al., 2021). In such settings, robustness is not merely a technical desideratum but a prerequisite for trust and reliability (Barreno et al., 2006).

The literature gap addressed by this article is thus twofold. First, there is a need for a unified theoretical narrative that situates time series adversarial research within the broader adversarial machine learning canon. Second, there is a need for critical reflection on the limitations of existing defenses, particularly in light of emerging attack strategies that challenge conventional assumptions about threat models and attacker capabilities (Apruzzese et al., 2023; Papernot et al., 2017). By engaging deeply with both foundational and contemporary sources, this article aims to contribute a durable scholarly resource for researchers, practitioners, and policymakers concerned with the security and reliability of deep learning systems.

The remainder of this article unfolds through an extensive methodological exposition, a detailed descriptive analysis of findings derived from the literature, and a wide-ranging discussion that situates these findings within ongoing theoretical and practical debates. Throughout, emphasis is placed on critical engagement, interpretive depth, and the articulation of future research trajectories informed by accumulated scholarly insight (Akhtar et al., 2021; Madry et al., 2017).

METHODOLOGY

The methodological orientation of this study is deliberately theoretical, interpretive, and synthesis-driven, reflecting the complex and heterogeneous nature of adversarial machine learning research across computer vision and time series domains. Rather than adopting an experimental or benchmark-centric methodology, which would necessarily privilege specific datasets, architectures, and threat models, this article employs a structured analytical methodology grounded in comprehensive literature integration, conceptual comparison, and critical interpretation. This approach is consistent with prior large-scale adversarial surveys and theoretical works that argue robustness cannot be fully understood through isolated empirical demonstrations alone, but instead requires contextualized reasoning across models, data modalities, and adversarial objectives (Akhtar et al., 2021; Joseph et al., 2018).

The first methodological pillar of this research is systematic literature consolidation. Foundational works on adversarial examples and robustness theory were examined to establish conceptual baselines regarding adversarial vulnerability, attacker knowledge assumptions, and defense taxonomies (Goodfellow et al., 2014; Szegedy et al., 2013; Madry et al., 2017). These foundational perspectives were then extended through engagement with domain-specific surveys and empirical studies focusing on computer vision and time series learning, enabling cross-domain conceptual mapping (Akhtar et al., 2021; Galib & Bashyal, 2023). The deliberate inclusion of both early and recent contributions allows for historical tracing of how adversarial research has evolved in response to emerging architectures, datasets, and real-world constraints.

The second methodological pillar involves comparative domain analysis. Computer vision and time series learning differ fundamentally in data structure, semantic interpretation, and preprocessing pipelines, yet they share underlying statistical learning principles that render them susceptible to adversarial manipulation (Ismail Fawaz et al., 2020; He et al., 2016). This study systematically contrasts how adversarial attacks are constructed, evaluated, and defended within each domain, focusing on conceptual mechanisms rather than numerical performance. For example, pixel-wise perturbations in images are analytically compared with temporal distortions, amplitude modulations, and frequency-domain manipulations in time series data, highlighting both parallels and divergences in adversarial strategy (Ding et al., 2023; Dong et al., 2023).

A third methodological component is defense taxonomy synthesis. Existing defenses are categorized not merely by implementation technique, but by their underlying epistemic assumptions about adversarial behavior. Adversarial training is examined as a robustness-by-exposure paradigm, defensive distillation as a smoothing-based approach, and ensemble defenses as diversity-driven robustness strategies (Madry et al., 2017; Papernot et al., 2016; Deng & Mu, 2023). These categories are critically analyzed across domains to assess their generalizability, scalability, and susceptibility to adaptive attackers, in line with critiques raised in both cybersecurity and machine learning literature (Apruzzese et al., 2023; Rosenberg et al., 2021).

The fourth methodological element is interpretive robustness evaluation. Traditional robustness evaluation often relies on worst-case accuracy metrics under specific attack algorithms, an approach that has been criticized for encouraging overfitting to known attacks and providing a false sense of security (Carlini & Wagner, 2017; Papernot et al., 2017). This study instead evaluates robustness claims through narrative synthesis of empirical findings reported across multiple studies, identifying recurring patterns of defense failure, attack transferability, and trade-off between accuracy and robustness (Akhtar et al., 2021; Machado et al., 2021). By aggregating interpretive evidence rather than numerical scores, the methodology mitigates the risk of narrow evaluation bias.

Methodological limitations are acknowledged explicitly. The reliance on published literature introduces potential publication bias, as unsuccessful defenses or negative results may be underreported (Pitropakis et al., 2019). Additionally, the absence of original experimental validation precludes direct performance comparison across methods. However, these limitations are offset by the study's breadth, depth, and theoretical integration, which collectively provide insights that are difficult to obtain through isolated experimental studies alone (Joseph et al., 2018).

Finally, ethical and epistemological considerations are integrated into the methodology. Adversarial machine learning research exists at the intersection of security and system design, raising questions about responsible

disclosure, dual-use knowledge, and the balance between robustness and transparency (Barreno et al., 2006; Hernandez-Castro et al., 2022). By foregrounding these considerations, the methodological framework situates technical analysis within a broader scholarly discourse on trustworthy artificial intelligence.

RESULTS

The results of this study emerge from a comprehensive synthesis of empirical findings, theoretical analyses, and comparative evaluations reported across the adversarial machine learning literature. Rather than presenting quantitative outcomes, the results are articulated as descriptive patterns and interpretive observations that collectively characterize the current state of adversarial robustness research in computer vision and time series learning (Akhtar et al., 2021; Galib & Bashyal, 2023).

A first major result concerns the universality of adversarial vulnerability. Across both computer vision and time series domains, deep learning models consistently exhibit susceptibility to carefully crafted perturbations, regardless of architecture complexity or training dataset size (Goodfellow et al., 2014; Ding et al., 2023). Residual networks, inception-style architectures, and ensemble classifiers all demonstrate failure modes under adversarial pressure, suggesting that vulnerability is an intrinsic property of high-capacity function approximators trained under standard optimization regimes (He et al., 2016; Ismail Fawaz et al., 2020). This finding aligns with theoretical arguments that adversarial examples exploit regions of the input space that are underrepresented or entirely absent in training data (Madry et al., 2017).

A second result highlights the domain-specific manifestation of adversarial strategies. In computer vision, attacks predominantly exploit pixel-level gradients and perceptual invariances, producing perturbations that remain visually imperceptible yet semantically disruptive (Carlini & Wagner, 2017; Moosavi-Dezfooli et al., 2016). In contrast, time series attacks often manipulate temporal continuity, phase alignment, or frequency characteristics, sometimes yielding perturbations that are visually subtle in plots but operationally significant in downstream decision-making (Dong et al., 2023; Ding et al., 2023). Despite these differences, both domains exhibit high degrees of attack transferability, wherein adversarial examples crafted for one model generalize to others with similar architectures or training data (Papernot et al., 2017; Galib & Bashyal, 2023).

A third result pertains to black-box adversarial feasibility. Studies consistently demonstrate that effective attacks do not require full access to model parameters or gradients, undermining assumptions that obscurity or proprietary architectures provide meaningful security (Papernot et al., 2017; Ding et al., 2023). Query-based and surrogate-model approaches enable attackers to approximate decision boundaries sufficiently well to induce misclassification, a result that holds across both image-based and temporal models (Dong et al., 2023; Akhtar et al., 2021). This observation reinforces the argument that robustness must be achieved through principled defense design rather than reliance on restricted access.

A fourth result concerns the conditional effectiveness of defense mechanisms. Adversarial training remains the most consistently effective defense across domains, improving robustness against known attacks while incurring computational cost and accuracy trade-offs (Madry et al., 2017; Bai et al., 2021). Defensive distillation and input preprocessing techniques provide limited protection and are often circumvented by adaptive attacks, confirming critiques that such methods primarily obscure gradients rather than fundamentally altering decision boundaries (Papernot et al., 2016; Carlini & Wagner, 2017). Ensemble-based defenses show promise in increasing attack difficulty and reducing transferability, particularly when model diversity is carefully engineered (Deng & Mu, 2023; Machado et al., 2021).

A fifth result underscores evaluation fragility. Robustness claims are highly sensitive to attack choice, parameterization, and evaluation protocol, leading to inconsistent conclusions across studies (Carlini & Wagner, 2017; Apruzzese et al., 2023). This issue is particularly pronounced in time series research, where standardized adversarial benchmarks are still emerging and evaluation practices vary widely (Dau et al., 2019; Galib & Bashyal, 2023). As a result, reported robustness improvements must be interpreted cautiously and contextualized within specific threat models.

Collectively, these results depict a research landscape characterized by significant conceptual convergence but persistent practical challenges. While understanding of adversarial mechanisms has deepened, achieving reliable, domain-agnostic robustness remains an open problem (Akhtar et al., 2021; Joseph et al., 2018).

DISCUSSION

The findings synthesized in this study invite a deeper theoretical and philosophical examination of adversarial robustness, extending beyond technical countermeasures to interrogate the assumptions underpinning modern machine learning. One of the most salient insights

emerging from the literature is that adversarial vulnerability is not a marginal flaw that can be patched through incremental adjustments, but rather a structural consequence of how high-dimensional models learn from finite data (Goodfellow et al., 2014; Madry et al., 2017). This realization challenges prevailing narratives of continuous performance improvement and calls for a reassessment of what robustness can realistically mean in adversarially exposed environments.

From a theoretical standpoint, the persistence of adversarial examples across domains suggests that vulnerability is deeply linked to the geometry of learned decision boundaries. In both image and time series models, training objectives prioritize average-case performance, leaving worst-case regions of the input space insufficiently constrained (Szegedy et al., 2013; Galib & Bashyal, 2023). Adversarial attacks exploit these regions, revealing a fundamental tension between generalization and robustness that cannot be resolved through data augmentation alone (Iwana & Uchida, 2021). This tension has prompted debate over whether robustness should be treated as a first-class optimization objective, even at the expense of nominal accuracy (Madry et al., 2017).

The discussion also highlights important epistemic concerns. Adversarial examples expose a disconnect between human semantic understanding and model-internal representations, raising questions about interpretability and trust (Hernandez-Castro et al., 2022). In time series contexts, this disconnect is particularly problematic, as temporal patterns often encode causal or contextual information that is not easily captured by static perturbation metrics (Ko et al., 2005; Ding et al., 2023). Consequently, robustness evaluation must consider domain semantics and operational impact, rather than relying solely on abstract norm-based constraints.

Defense strategies are likewise subject to critical scrutiny. Adversarial training, while effective, embodies an arms-race dynamic in which defenses are continually challenged by stronger attacks (Bai et al., 2021; Apruzzese et al., 2023). Ensemble defenses introduce diversity as a robustness mechanism, but their success depends on careful design choices that are not yet theoretically well understood (Deng & Mu, 2023). Moreover, the computational and environmental costs of large-scale adversarial training raise practical concerns about scalability and sustainability, particularly in resource-constrained settings (Shaukat et al., 2020).

Another critical dimension concerns evaluation culture. The literature reveals a tendency toward attack-specific robustness claims, which can mislead stakeholders about real-world security (Carlini & Wagner, 2017). Bridging the gap between adversarial research and practice requires threat models grounded in realistic attacker capabilities and incentives, a point emphasized by recent critiques of gradient-centric assumptions (Apruzzese et al., 2023). For time series applications in finance, healthcare, and industrial systems, adversaries may operate under constraints that differ markedly from those assumed in academic benchmarks (Corizzo & Rosen, 2024; Anthi et al., 2021).

Future research directions emerge naturally from these discussions. There is a pressing need for domain-aware robustness frameworks that integrate statistical learning theory with contextual knowledge of data generation processes (Akhtar et al., 2021). In time series learning, this may involve leveraging temporal causality, multi-scale representations, and domain-specific invariances as robustness anchors (Hewage et al., 2021). Cross-domain collaboration between computer vision and time series researchers could accelerate the transfer of insights and foster more unified robustness paradigms.

Ultimately, adversarial machine learning compels a reconsideration of what it means for a model to "understand" data. Robustness, in this sense, is not merely resistance to perturbation, but alignment between model behavior and human-intended semantics under a wide range of conditions (Joseph et al., 2018; Papernot et al., 2016). Achieving such alignment remains one of the central intellectual challenges of contemporary artificial intelligence research.

CONCLUSION

This article has presented an extensive, integrative examination of adversarial attacks and defenses across computer vision and time series learning, emphasizing theoretical depth, cross-domain synthesis, and critical interpretation. The analysis demonstrates that adversarial vulnerability is a pervasive and structurally rooted phenomenon, manifesting consistently across data modalities and model architectures. While significant advances have been made in understanding and mitigating adversarial threats, current defenses remain constrained by evaluation fragility, computational cost, and adaptive attacker dynamics. By situating time series adversarial research within the broader adversarial machine learning discourse, this study contributes a unified conceptual framework that clarifies shared challenges and highlights domain-specific nuances. Future progress will depend on moving beyond isolated technical fixes toward robustness paradigms that integrate theory, domain knowledge, and realistic threat modeling, thereby advancing the

development of trustworthy and resilient intelligent systems.

REFERENCES

1. Barreno, M., Nelson, B., Sears, R., Joseph, A. D., Tygar, J. D. Can machine learning be secure? Proceedings of the ACM Symposium on Information, Computer and Communications Security, 2006.

2. Akhtar, N., Mian, A., Kardan, N., Shah, M. Advances in adversarial attacks and defenses in computer vision: A survey. IEEE Access, 2021.

3. Ding, D., Zhang, M., Feng, F., Huang, Y., Jiang, E., Yang, M. Black-box adversarial attack on time series classification. Proceedings of the AAAI Conference on Artificial Intelligence, 2023.

4. Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. IEEE Symposium on Security and Privacy, 2016.

5. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A. Towards deep learning models resistant to adversarial attacks. arXiv preprint, 2017.

6. Goodfellow, I. J., Shlens, J., Szegedy, C. Explaining and harnessing adversarial examples. arXiv preprint, 2014.

7. Carlini, N., Wagner, D. Towards evaluating the robustness of neural networks. IEEE Symposium on Security and Privacy, 2017.

8. Dau, H. A., Bagnall, A., Kamgar, K., Yeh, C. C. M., Zhu, Y., Gharghabi, S., Keogh, E. The UCR time series archive. IEEE/CAA Journal of Automatica Sinica, 2019.

9. Galib, A. H., Bashyal, B. On the susceptibility and robustness of time series models through adversarial attack and defense. arXiv preprint, 2023.

10. Deng, Y., Mu, T. Understanding and improving ensemble adversarial defense. Advances in Neural Information Processing Systems, 2023.

11. Corizzo, R., Rosen, J. Stock market prediction with time series data and news headlines: a stacking ensemble approach. Journal of Intelligent Information Systems, 2024.

12. Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Petitjean, F. InceptionTime: Finding AlexNet for time series classification. Data Mining and Knowledge Discovery, 2020.

13. He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

14. Apruzzese, G., Anderson, H. S., Dambra, S., Freeman, D., Pierazzi, F., Roundy, K. Real attackers don't compute gradients. IEEE Conference on Secure and Trustworthy Machine Learning, 2023.

15. Joseph, A. D., Nelson, B., Rubinstein, B. I., Tygar, J. Adversarial machine learning. Cambridge University Press, 2018.

16. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., Swami, A. The limitations of deep learning in adversarial settings. IEEE European Symposium on Security and Privacy, 2016.

17. Machado, G. R., Silva, E., Goldschmidt, R. R. Adversarial machine learning in image classification: A survey toward the defender's perspective. ACM Computing Surveys, 2021.

18. Ko, M. H., West, G., Venkatesh, S., Kumar, M. Online context recognition in multisensor systems using dynamic time warping. IEEE International Conference on Intelligent Sensors, Sensor Networks and Information Processing, 2005.

19. Hewage, P., Trovati, M., Pereira, E., Behera, A. Deep learning-based effective fine-grained weather forecasting model. Pattern Analysis and Applications, 2021.